链滴

JAVA 通过 epublib 解析 EPUB 格式的电子书

作者：hjljy
原文链接：https://ld246.com/article/1583655592242
来源网站：链滴

什么是 epub 格式

就像视频文件有 MP4,AVI,RMVB 等等一样！电子书也有很多种格式：一文看懂mobi,azw3,epub格电子书

可以将 epub 格式的电子书更换后缀名，然后解压打开查看里面的文件信息。

Java 解析 Epub 格式电子书

刚接到这个需求的时候，在网上找了很久，没找到很好的解析方法，最后找到了 epublib 这个解析库但是下载对应的 jar 很麻烦，最终在 maven 仓库搜索找到了。

epublib 解析库

epublib：a Java library for reading and writing epub files　　(一个用于读写 epub 文件的 Java)

GitHub：https://github.com/psiegman/epublib

官方网址：http://www.siegmann.nl/epublib

API 地址：http://www.siegmann.nl/static/epublib/apidocs/（英文的）

第一步：引入对应的 pom 文件

```
<dependency>
    <groupId>com.positiondev.epublib</groupId>
    <artifactId>epublib-core</artifactId>
    <version>3.1</version>
```

```
</dependency>
<!--html解析 -->
<dependency>
    <groupId>org.jsoup</groupId>
    <artifactId>jsoup</artifactId>
    <version>1.12.1</version>
</dependency>
```

第二步：常用关键类

1.Book 表示电子书。通过 book 对象可以获取 resource，Metadata 等具体内容
2.Resource 表示电子书内容资源， 一个 Resource 就是电子书的一部分内容，这资源信息可以 html,css,js,图片等；
3.Resources 表示电子书全部的 Resource 对象。可以用过 id,herf,MediaType 来获取对应的 Resource 对象
4.MetaData 表示电子书的开篇信息。比如，作者，出版社，语言等；
5.Spine 电子书的 resource 顺序，有人说是目录信息，其实不是，是 resource 的阅读顺序，线性结构的
6.TableOfContent 电子书的目录信息，是树形结构的。可以获取到目录对应的resource。
7.MediaType Resource 的类型描述。用于说明此 Resource 是何种类型（CSS/JS/图片/HTML/VEDIO 等）。

第三步：解析一个epub文件

```
public static void main(String[] args) {

    File file = new File("E:\\Download\\红楼梦.epub");
    InputStream in = null;
    try {
        //从输入流当中读取epub格式文件
        EpubReader reader = new EpubReader();
        in = new FileInputStream(file);
        Book book = reader.readEpub(in);
        //获取到书本的头部信息
        Metadata metadata = book.getMetadata();
        System.out.println("FirstTitle为："+metadata.getFirstTitle());
        //获取到书本的全部资源
        Resources resources = book.getResources();
        System.out.println("所有资源数量为："+resources.size());
        //获取所有的资源数据
        Collection<String> allHrefs = resources.getAllHrefs();
        for (String href : allHrefs) {
            Resource resource = resources.getByHref(href);
            //data就是资源的内容数据，可能是css,html,图片等等
            byte[] data = resource.getData();
            // 获取到内容的类型  css,html,还是图片
            MediaType mediaType = resource.getMediaType();
        }
        //获取到书本的内容资源
        List<Resource> contents = book.getContents();
        System.out.println("内容资源数量为："+contents.size());
        //获取到书本的spine资源 线性排序
        Spine spine = book.getSpine();
```

```
            System.out.println("spine资源数量为："+spine.size());
            //通过spine获取所有的数据
            List<SpineReference> spineReferences = spine.getSpineReferences();
            for (SpineReference spineReference : spineReferences) {
                Resource resource = spineReference.getResource();
                //data就是资源的内容数据，可能是css,html,图片等等
                byte[] data = resource.getData();
                // 获取到内容的类型  css,html,还是图片
                MediaType mediaType = resource.getMediaType();
            }
            //获取到书本的目录资源
            TableOfContents tableOfContents = book.getTableOfContents();
            System.out.println("目录资源数量为："+tableOfContents.size());
            //获取到目录对应的资源数据
            List<TOCReference> tocReferences = tableOfContents.getTocReferences();
            for (TOCReference tocReference : tocReferences) {
                Resource resource = tocReference.getResource();
                //data就是资源的内容数据，可能是css,html,图片等等
                byte[] data = resource.getData();
                // 获取到内容的类型  css,html,还是图片
                MediaType mediaType = resource.getMediaType();
                if(tocReference.getChildren().size()>0){
                    //获取子目录的内容
                }
            }
        } catch (Exception e) {
            e.printStackTrace();
        } finally {
            //一定要关闭资源
            try {
                if (in != null) {
                    in.close();
                }
            } catch (IOException e) {
                e.printStackTrace();
            }
        }
    }
```

注意事项

1 解析后得到的data内容数据是html格式的富文本内容，如果需要纯文本，可以通过jsoup获取P标的文本内容就可以了，但是获取后的纯文本排版就会乱。

2 资源当中可能会存在图片和css等等，不在目录或者spine当中的内容,可以通过Resources.getByHre等方法获取。